GlusterFS and RHS for SysAdmins

An In-Depth Look with Demos

Niels de Vos
Sr. Software Maintenance Engineer
Red Hat Global Support Services

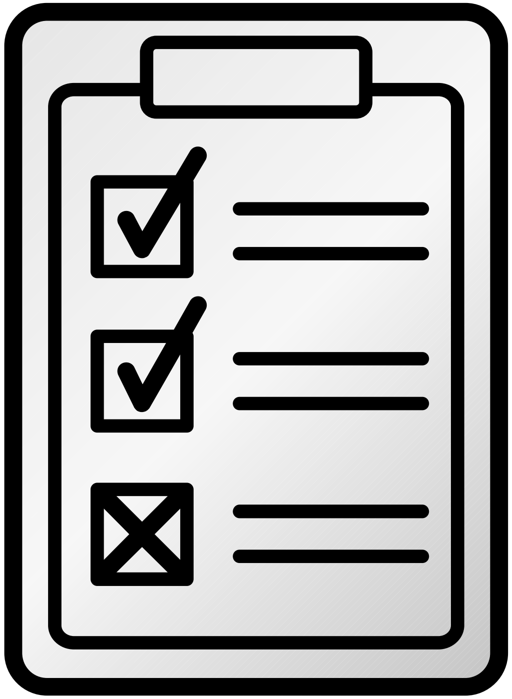
FISL – 7 May 2014

Introduction

- Name: Niels de Vos
- Company: Red Hat
- Department: Global Support Services
- Job title: Sr. Software Maintenance Engineer
- Duties:
  - assist with solving complex customer support cases, write bugfixes/patches, document solutions
  - Sub-maintainer for Gluster/NFS, release-maintainer for glusterfs-3.5 (current stable version)

Agenda

- Technology Overview & Use Cases
- Technology Stack
- Under the Hood
- Volumes and Layered Functionality
- Data Access



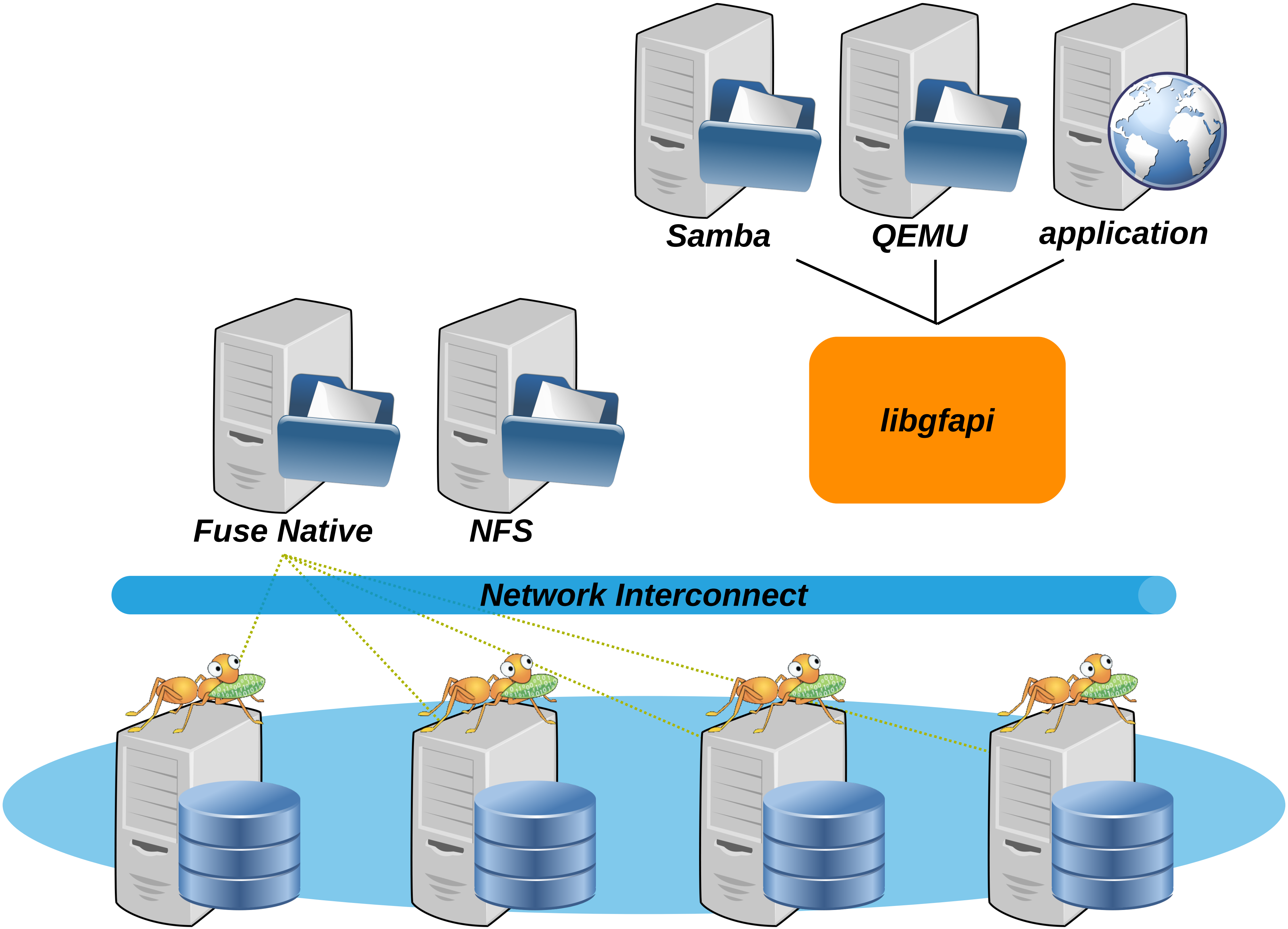
Technology Overview

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

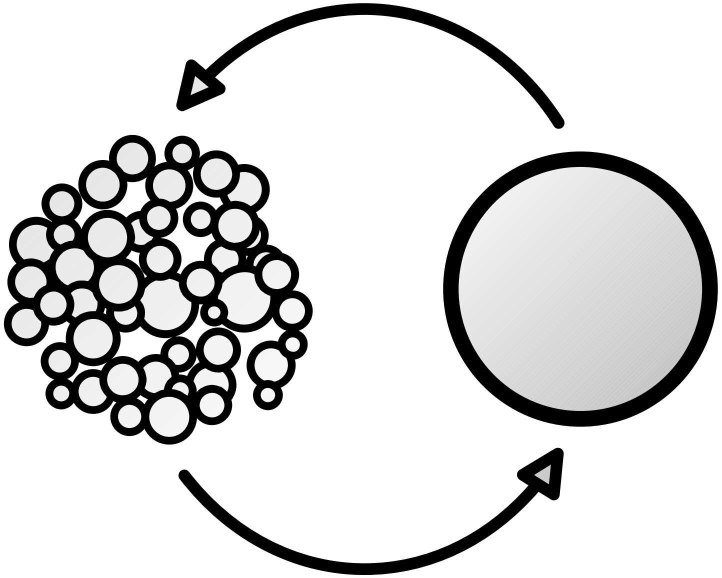
What is GlusterFS?

- Clustered Scale-out **General Purpose** Storage Platform
    - POSIX-y Distributed File System
    - ...and so much more
- Built on Commodity systems
    - x86_64 Linux
    - POSIX filesystems underneath (XFS, EXT4)
- No Metadata Server
- Standards-Based – Clients, Applications, Networks
- Modular Architecture for Scale and Functionality


Samba
QEMU
application
libgfapi
Fuse Native
NFS
Network Interconnect

What is Red Hat Storage?

- Enterprise Implementation of GlusterFS
- Integrated Software Appliance
  - RHEL + XFS + GlusterFS
- Certified Hardware Compatibility
- Subscription Model
- 24x7 Premium Support



GlusterFS vs. Traditional Solutions

- A basic NAS has limited scalability and redundancy
- Other distributed filesystems are limited by metadata service
- SAN is costly & complicated, but high performance & scalable
- *GlusterFS is...*
  - *Linear Scaling*
  - *Minimal Overhead*
  - *High Redundancy*
  - *Simple and Inexpensive Deployment*

Use Cases

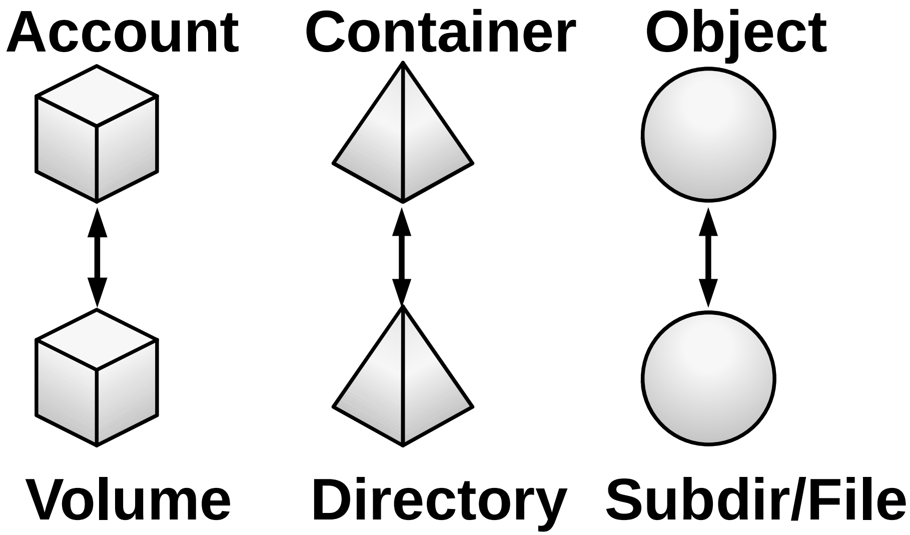
Common Solutions

- Large Scale File Server
- Media / Content Distribution Network (CDN)
- Backup / Archive / Disaster Recovery (DR)
- High Performance Computing (HPC)
- Infrastructure as a Service (IaaS) storage layer
- Database offload (blobs)
- Unified Object Store + File Access

Hadoop – Map Reduce

- Access data within and outside of Hadoop
- No HDFS name node single point of failure / bottleneck
- Seamless replacement for HDFS
- Scales with the massive growth of big data

Swift



Account

Container

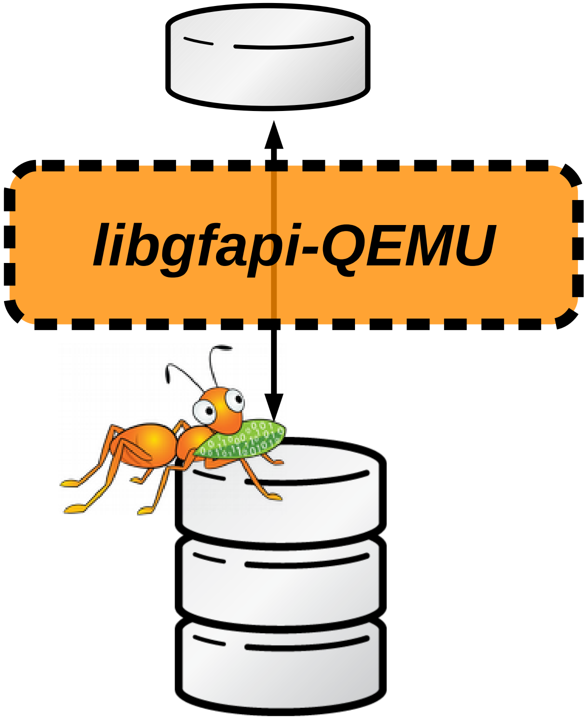
Object

Volume

Directory

Subdir/File

Cinder / Glance



libgfapi-QEMU

Technology Stack

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

Terminology

- Brick
  - Fundamentally, a filesystem mountpoint
  - A unit of storage used as a ***capacity*** building block
- Translator
  - Logic between the file bits and the Global Namespace
  - Layered to provide GlusterFS ***functionality***

Everything is Modular

Terminology

- Volume
  - Bricks combined and passed through translators
  - Ultimately, what's presented to the end user
- Peer / Node
  - Server hosting the brick filesystems
  - Runs the Gluster daemons and participates in volumes

Disk, LVM, and Filesystems

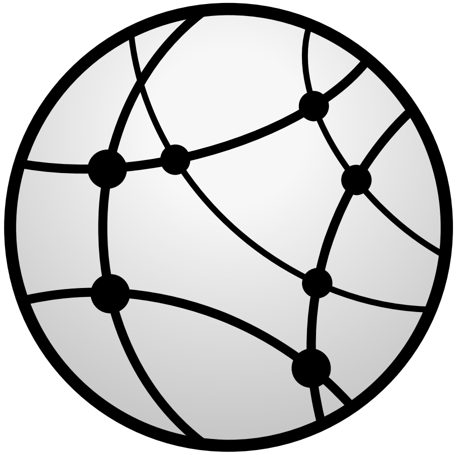
- Direct-Attached Storage (DAS)

  -or-

- Just a Bunch Of Disks (JBOD)
- Hardware RAID
  - RHS: RAID 6 required
- Logical Volume Management (LVM)
- POSIX filesystem w/ Extended Attributes (EXT4, XFS, BTRFS, ...)
  - RHS: XFS required

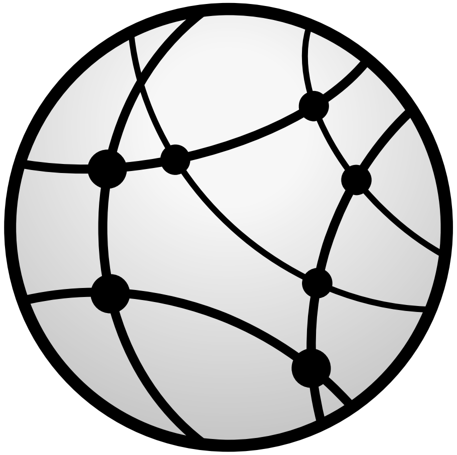
Data Access Overview

- GlusterFS Native Client
  - Filesystem in Userspace (FUSE)
- NFS
  - Built-in Service
- SMB/CIFS
  - Samba server required;
    NOW libgfapi-integrated!



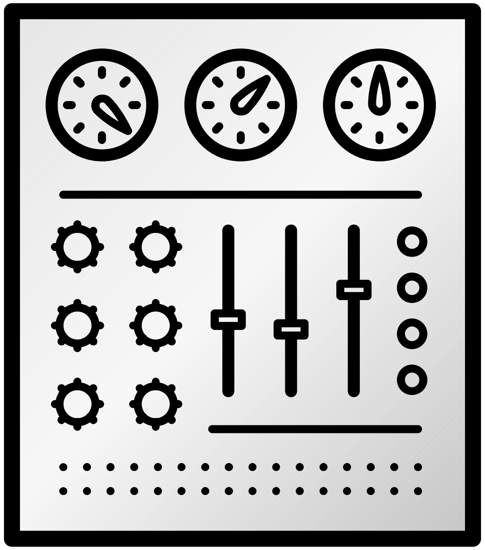
Data Access Overview

- Gluster For OpenStack (G4O; aka UFO)
  - Simultaneous object-based access via OpenStack Swift
- NEW! libgfapi flexible abstracted storage
  - Integrated with upstream Samba and NFS-Ganesha



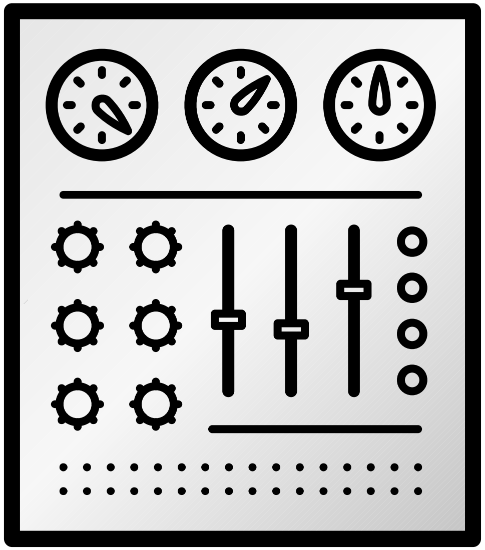
Gluster Components

- `glusterd`
  - Management daemon
  - One instance on each GlusterFS server
  - Interfaced through `gluster` CLI
- `glusterfsd`
  - GlusterFS brick daemon
  - One process for each brick on each server
  - Managed by `glusterd`



Gluster Components

- `glusterfs`
  - Volume service daemon
  - One process for each volume service
    - NFS server, FUSE client, Self-Heal, Quota, .
- `mount.glusterfs`
  - FUSE native client mount extension
- `gluster`
  - Gluster Console Manager (CLI)

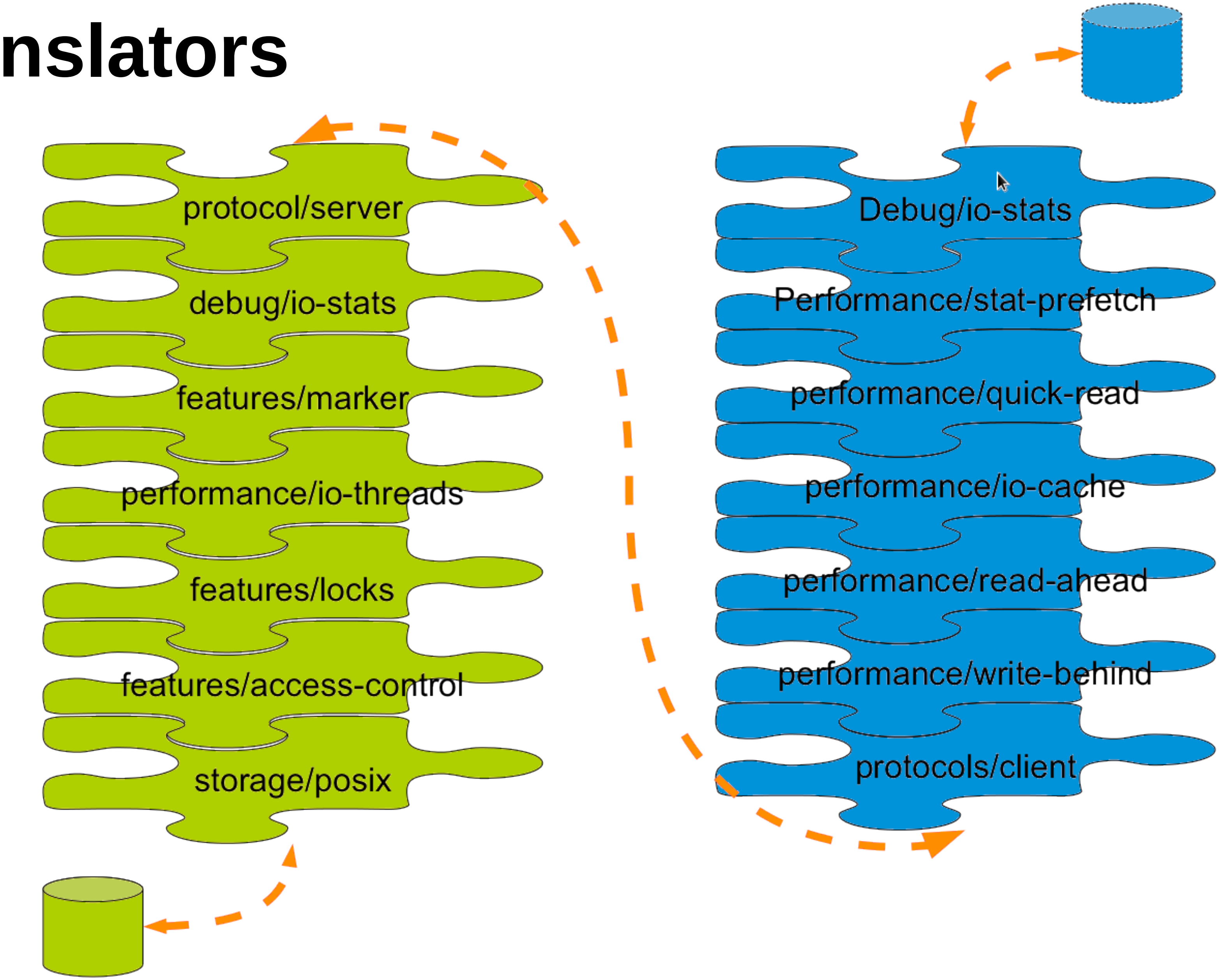


Under the Hood

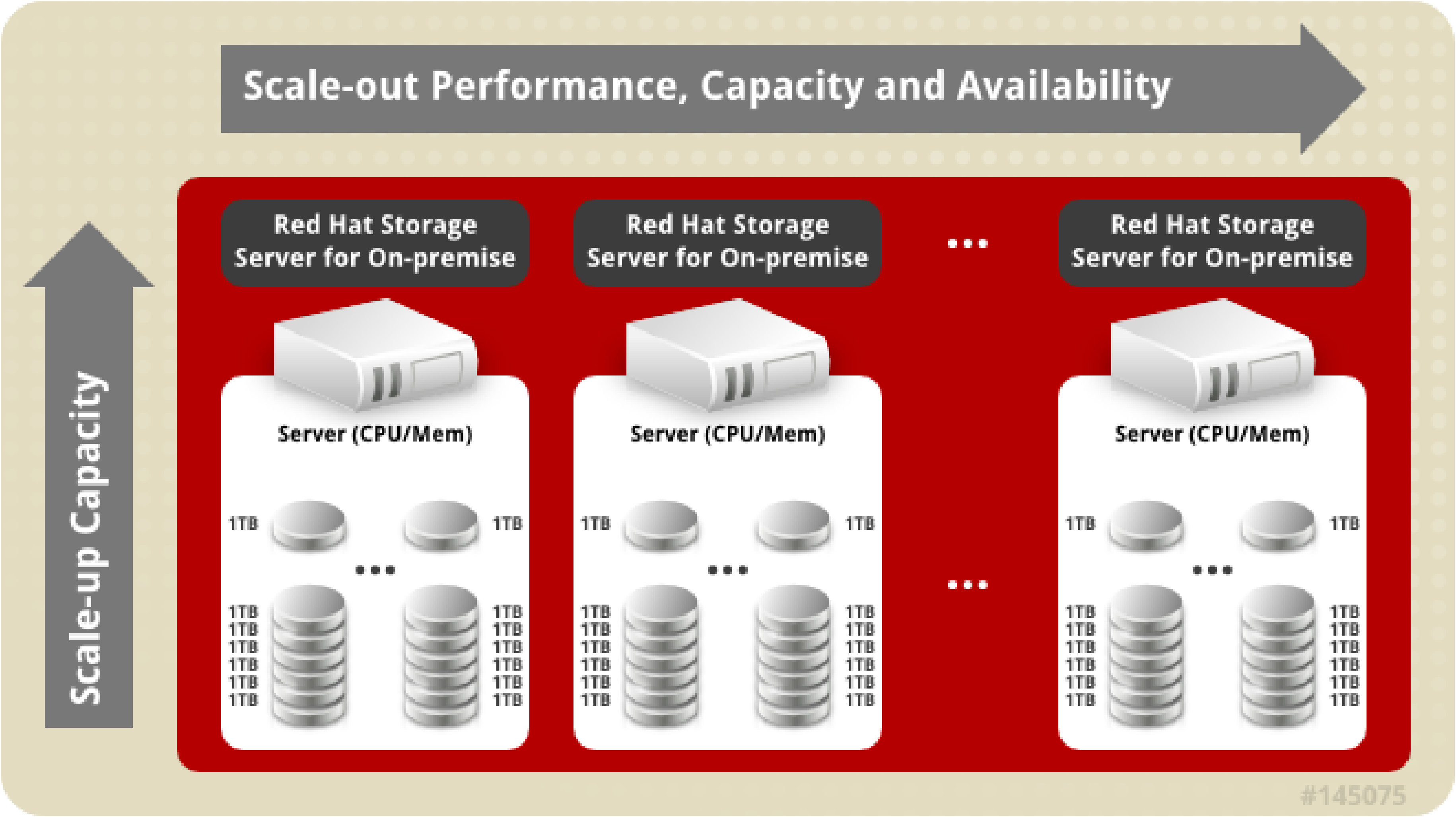
GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

Translators



Up and Out!


Scale-out Performance, Capacity and Availability
Scale-up Capacity
Red Hat Storage Server for On-premise
Server (CPU/Mem)
1TB
...
Red Hat Storage Server for On-premise
Server (CPU/Mem)
...
Red Hat Storage Server for On-premise
Server (CPU/Mem)
#145075

Your Storage Servers are Sacred!

- Don't touch the brick filesystems directly!
- They're Linux servers, but treat them like appliances
  - Separate security protocols
  - Separate access standards
- Don't let your Jr. Linux admins in!
  - A well-meaning sysadmin can quickly break your system or destroy your data

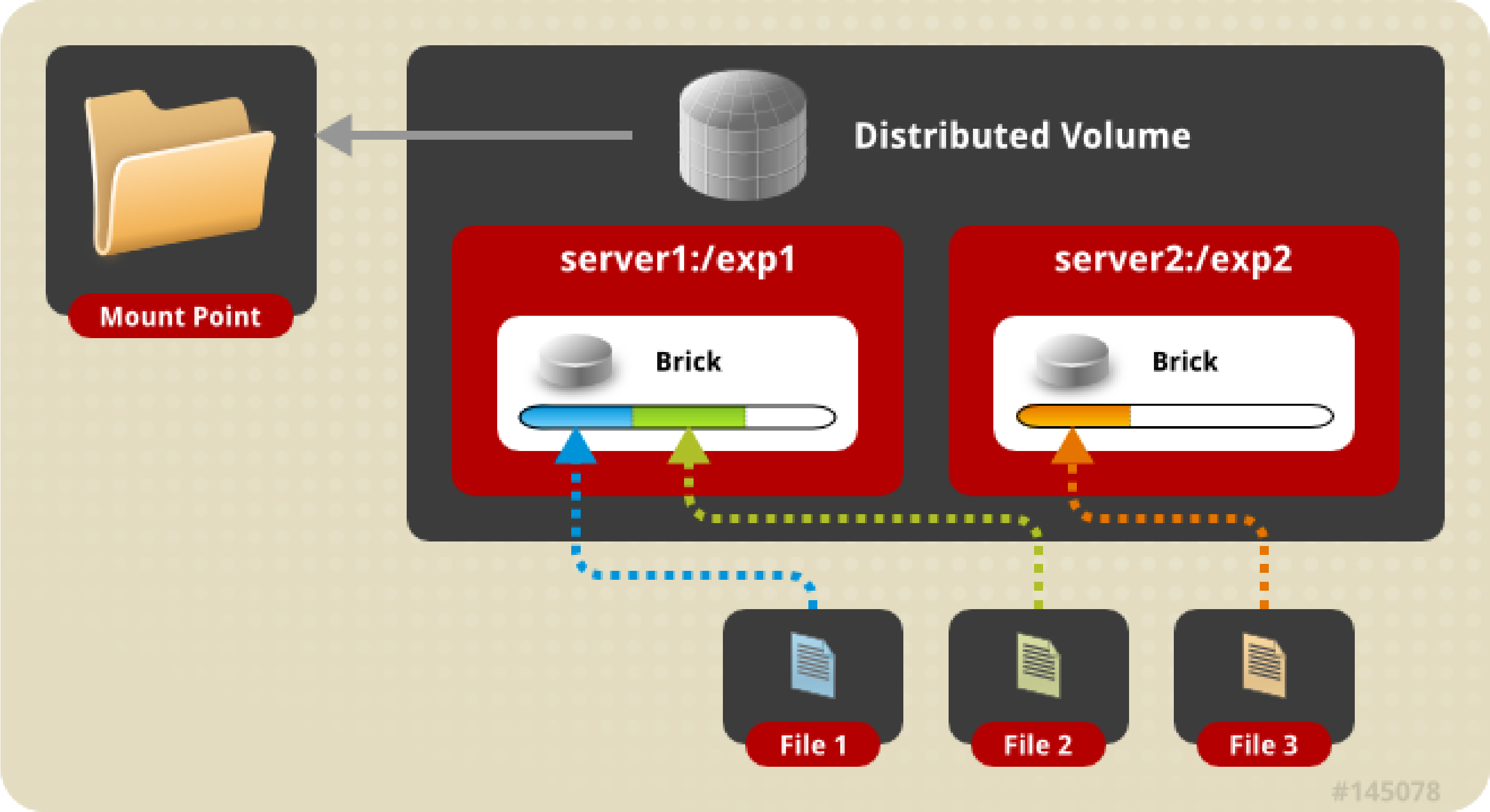
Basic Volumes

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

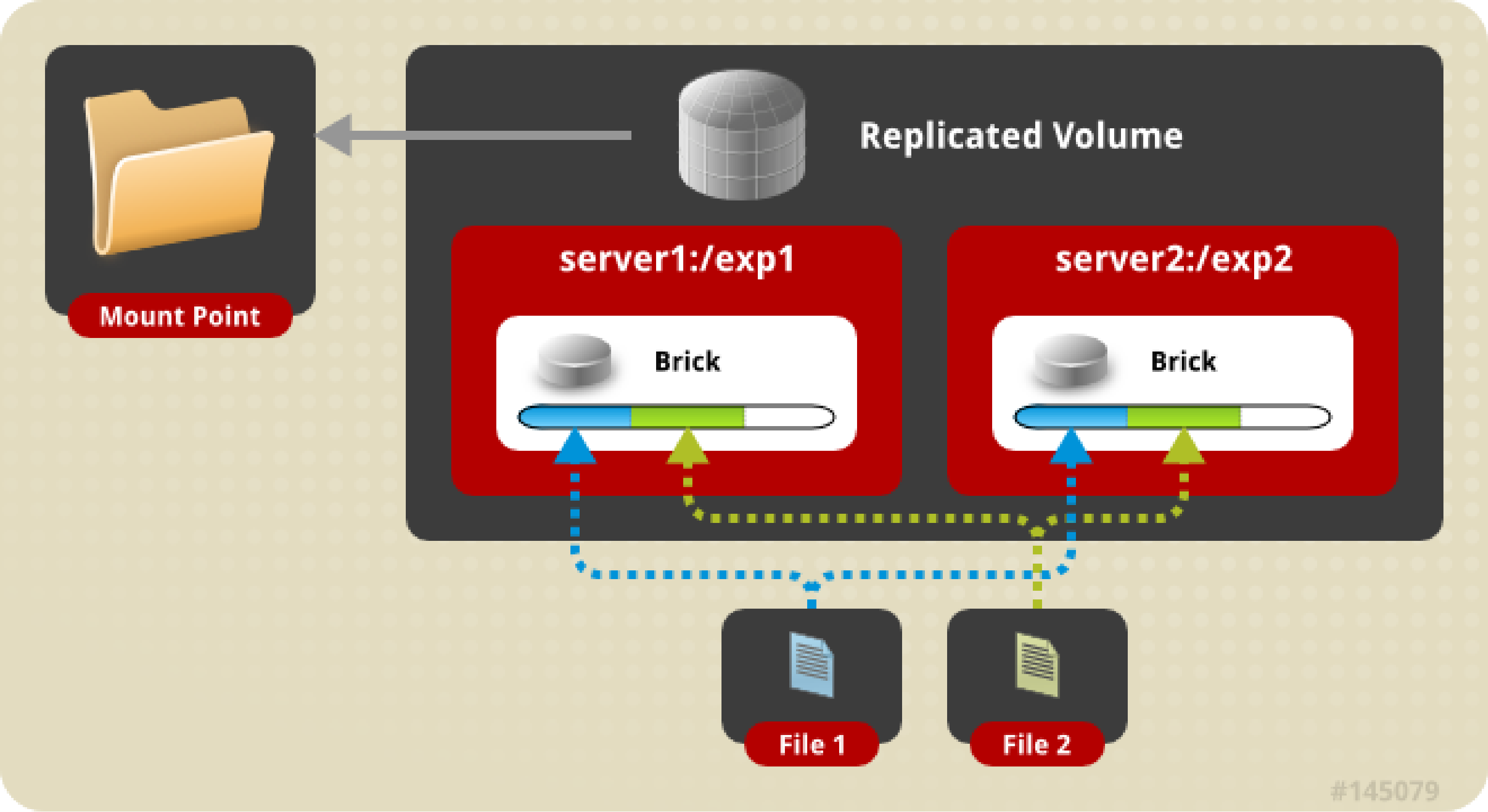
Distributed Volume

- Files “evenly” spread across bricks
- *Similar* to file-level RAID 0
- Server/Disk failure could be catastrophic



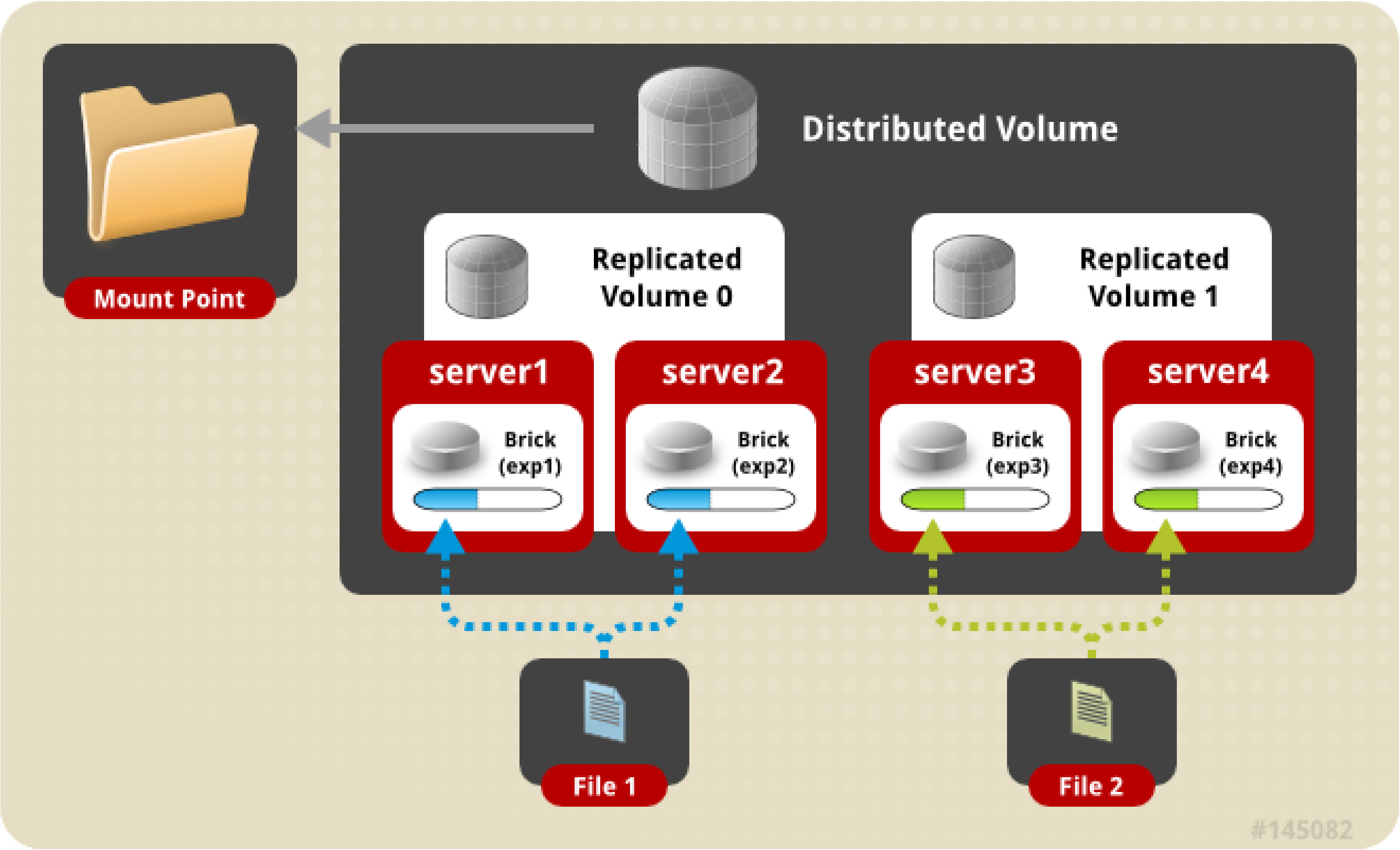
Replicated Volume

- Copies files to multiple bricks
- *Similar* to file-level RAID 1



Distributed Replicated Volume

- Distributes files across replicated bricks



Data access

GlusterFS and RHS for the SysAdmin

An In-Depth Look and Demo

GlusterFS Native Client (FUSE)

- FUSE kernel module allows the filesystem to be built and operated entirely in userspace
- Specify mount to any GlusterFS server
- Native Client fetches volfile from mount server, then communicates directly with all nodes to access data
- Recommended for high concurrency and high write performance
- Load is inherently balanced across distributed volumes

NFS

- Standard NFS v3 clients
- Standard automounter is supported
- Mount to any server, or use a load balancer
- GlusterFS NFS server includes Network Lock Manager (NLM) to synchronize locks across clients
- Better performance for reading many small files from a single client
- Load balancing must be managed externally

NEW! libgfapi

- Introduced with GlusterFS 3.4
- User-space library for accessing data in GlusterFS
- Filesystem-like API
- Runs in application process
- no FUSE, no copies, no context switches
- ...but same volfiles, translators, etc.

SMB/CIFS

- NEW! In GlusterFS 3.4 – Samba + libgfapi
  - No need for local native client mount & re-export
  - Significant performance improvements with FUSE removed from the equation
- Must be setup on each server you wish to connect to via CIFS
- CTDB is required for Samba clustering

Do it!

- Build a test environment in VMs in just minutes!
- Get the bits:
  - Fedora 20 has GlusterFS packages natively
  - RHS 2.1 ISO available on Red Hat Portal
  - Go upstream: www.gluster.org
  - Amazon Web Services (AWS)
    - Amazon Linux AMI includes GlusterFS packages
    - RHS AMI is available

Thank You!

Slides Available at: **http://people.redhat.com/ndevos/talks/fisl15**

- **ndevos@redhat.com**

  storage-sales@redhat.com
- **RHS:**

  www.redhat.com/storage
- **GlusterFS:**

  www.gluster.org
- **Red Hat Global Support Services:** **access.redhat.com/support**

@Glusterorg

@RedHatStorage

Gluster

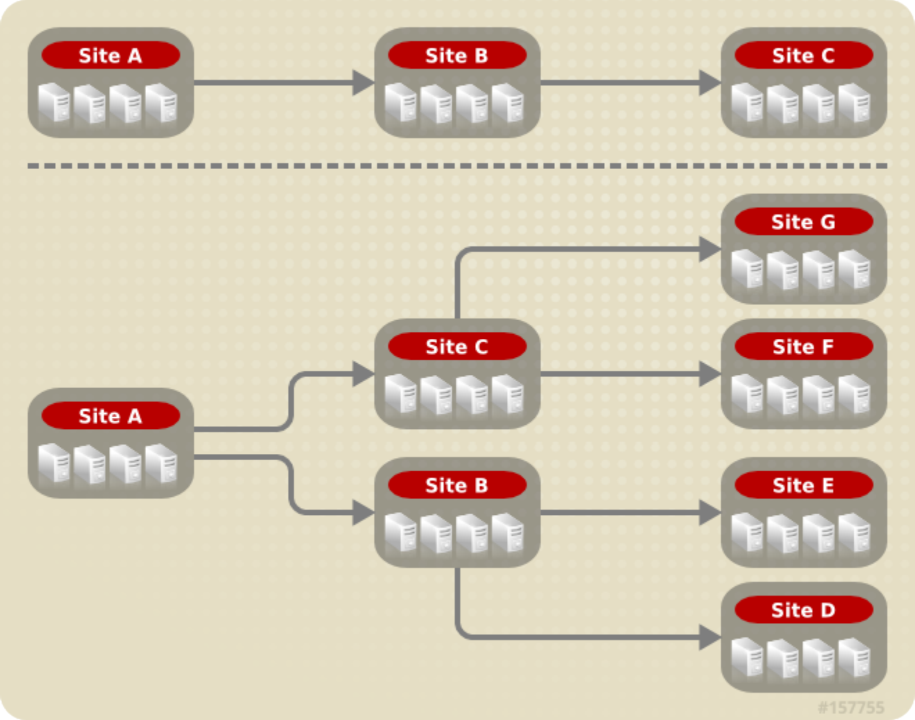
Red Hat Storage

GlusterFS and RHS for SysAdmins

An In-Depth Look with Demos

Geo Replication

- Asynchronous across LAN, WAN, or Internet
- Master-Slave model -- Cascading possible
- Continuous and incremental
- Data is passed between defined master and slave **only**



Elastic Hash Algorithm

- No central metadata
  - No Performance Bottleneck
  - Eliminates risk scenarios
- Location hashed intelligently on filename
  - Unique identifiers, similar to md5sum
- The “Elastic” Part
  - Files assigned to virtual volumes
  - Virtual volumes assigned to multiple bricks
  - Volumes easily reassigned on the fly